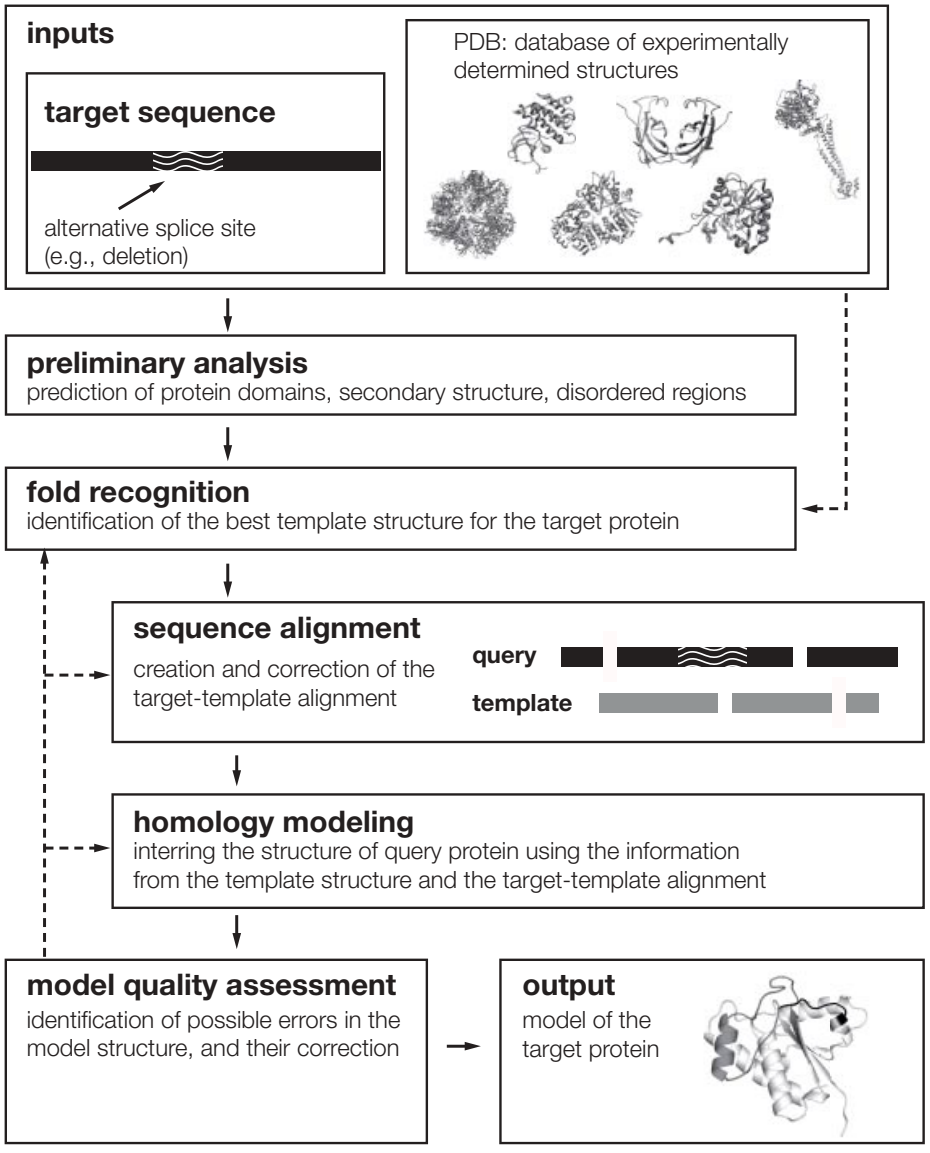
Structure Prediction for Alternatively Spliced Proteins



Outcome:
Model of the alternatively spliced protein.

Question answered:
What is the closest template structure for the target protein?
How to align target and template sequences.
How to assess the quality of the structural model.
How to interpret structural modifications resulting from splicing.

54
Structure Prediction for Alternatively Spliced Proteins

Lukasz Kozlowski, Jerzy Orlowski, and Janusz M. Bujnicki

Abstract

The abundance of different protein variants generated by alternative splicing (AS) is not fully represented in the Protein Structure Database (PDB). Among the already limited set of proteins with experimentally determined structures deposited in PDB, only a small fraction have structures of more than one splicing variant. In the absence of experimental data, computational methods can be used to predict the structure of protein isoforms. In this chapter, the established approaches to computational protein structure prediction will be briefly described, and their use for studying differences that result from AS will be detailed. In particular, it will be noted when a reliable structural model can be obtained for a given protein sequence. Topics of finding the best template structures for modeling individual protein domains, predicting the structure of long, multidomain proteins, and assessing the quality of theoretical models, together with error correction, will also be addressed. The suggested protocol will be illustrated by the structure prediction of isoform C of phosphotyrosine phosphatase (LMPTP-C).

54.1
Theoretical Background

In 1961, based on studies with ribonuclease A, Christian Anfinsen hypothesized that the structure of a protein in its native environment would be determined by the protein's amino acid sequence, and that this native conformation is the one in which the Gibbs free energy of the system is lowest [1]. Since then, the prediction of a protein's structure from its amino acid sequence has become the "holy grail" of computational biology. However, despite great efforts having been made, a universal algorithm with which to infer protein structure has not yet been developed. Nonetheless, several methods have been developed that can provide useful models, depending on a variety of conditions.

There are two major approaches for protein structure prediction. The first approach, termed "comparative modeling" or "template-based modeling," is based on the experimental observation that evolutionarily related proteins usually retain similar structures, despite an accumulation of substitutions at the level of amino acid sequence, and that the structure changes very slowly compared to the sequence [2]. Thus, an experimentally determined structure of one protein can be used as a template to model the structure of another related protein (a modeling target) by simulating the process of evolution at the sequence level. Modification of the template to "transmute" it into the target requires only limited computation. Therefore, comparative modeling does not require special computer resources, and can be easily carried out on a personal workstation using computer programs that are simple to install and use.

Template-based modeling requires that, for a given target sequence, a structurally similar template is identified, and a correct target–template sequence alignment is

Alternative pre-mRNA Splicing: Theory and Protocols, First Edition. Edited by Stefan Stamm, Chris Smith, and Reinhard Lührmann.

determined. This procedure can generate a model only for such region of the target sequence for which a structurally characterized template exists and can be detected. For multidomain proteins, the structures of individual domains must normally be modeled separately.

Template-based modeling relies mostly on copying such elements of the template that are inferred to be essentially the same in the target (e.g., backbone conformation in aligned regions, and side-chain conformation of invariant amino acid residues), while all modifications (residue substitutions, insertions or deletions) are introduced in such a way as to minimize the disruption of the conserved core. The more similar the template sequence is to the target, the easier it is to identify in the database and the fewer errors are introduced at the stage of target–template alignment. Consequently, the likelihood of success of template-based modeling and the quality of the model (i.e., its similarity to the real structure) depends largely on the evolutionary distance between the target and the template. Comparative modeling based on templates with $>$50% sequence identity to the target can yield structures of accuracy comparable with medium-resolution crystallography or NMR. However, as sequence similarity decreases, so too does the structural divergence, and models based on remotely related templates typically exhibit deviations from true structures, in particular in highly variable regions such as loops. Another problem is that with increased evolutionary distance, errors at the level of sequence alignment become more likely, and usually their detection and correction requires special expertise. It must be emphasized that template-based prediction does not take into account the possibility of global structural changes such as domain swapping, which may occur as a consequence of events such as mutation or alternative splicing (AS). Essentially, template-based modeling implies that the target is modeled in the same structural functional state as the template structure (e.g., with or without ligands, with open or closed conformation, etc.). Template-based modeling is, therefore, not an appropriate tool with which to model conformational changes.The procedure of comparative modeling comprises three steps:

1) Identification of the template structure and generation of the target–template alignment to establish which amino acid residues of the target correspond to which residues of the template. Recommended program: GENESILICO METASERVER [3] (developed in the laboratory of the authors).
2) Construction of a three-dimensional (3-D) model of the target, based on structural information from the template. Recommended programs: MODELLER [4] or SWISS-MODEL [5].
3) Assessment of the global and local quality of the model and correction of potential errors at the level of alignment (Step 1) and/or model construction (Step 2). Recommended programs: PROQ [6] or METAMQAP [7] (developed in the laboratory of the authors).

Whilst a detailed description of these steps is beyond the scope of this protocol, an example protocol will be described in Section 54.3.

Recent analyses [23] have estimated that automated procedures of comparative modeling can provide structural models for about 70% of proteins sequences present in public databases (for this fraction of sequences, at least one domain can be modeled). For about 30% of sequences no template structures can be reliably detected by fully automated methods. In many cases remotely related templates do exist, but their detection requires expert knowledge and the use of nonstandard tools. For protein sequences without templates, “template-free” structure prediction methods have been developed that sample a large number of alternative conformations and attempt to identify the one with the lowest Gibbs free energy, following Anfinsen’s hypothesis.

Compared to template-based methods, template-free modeling is computationally very time-consuming as it requires complicated calculations to be made for multiple conformations. Even with modern supercomputers, it is extremely difficult to simulate “ab initio” more than a few microseconds of the physical process of

folding for very small proteins. Therefore, knowledge-based "*de-novo*" methods have been developed that restrict the search to conformations similar to those observed in known protein structures, and that replace the calculation of physical energies with much simpler scoring functions. Although such methods are capable of generating conformations that are close to the native structure, their scoring functions are inaccurate and they cannot guarantee the identification of a correct solution. An example of a "*de novo*" method that can be installed and run on a personal workstation is ROSETTA [8], which assembles models from short fragments derived from previously determined protein structures. However, as the number of possible conformations increases rapidly with the protein length, template-free modeling on a personal workstation (e.g., with ROSETTA) is still practically limited to protein domains of less than 80 residues. The template-free modeling of larger proteins requires the parallelization of calculations and, for example, the use of computing clusters. However, it must be emphasized that the likelihood of obtaining native-like template-free models for sequences longer than 100 residues is currently quite low, regardless of the method or computing power used, due to limited sampling and inaccuracies of the scoring functions.

54.2 Protocol

The structure prediction of alternatively spliced protein variants using template-based modeling is carried out in step-wise fashion as follows.

54.2.1 Primary Structure Analysis

Most modeling methods have been developed to deal with individual domains only. As many proteins (in particular those from Eukaryota) consist of multiple domains, the actual modeling should be preceded by a determination of whether the target sequence comprises one or more domains, whether these domains are likely to fold into globular structures, by an analysis of the relationships of these domains to other protein sequences, and by the determination of the best modeling approach for each domain (template-based or template-free). As the first step, it is recommended to compare the target sequence to a database of protein families and/or domains, such as PFAM [24].

54.2.2 Predicting Disordered Regions

Some proteins or their parts have no stable structure in solution, and they fluctuate between different conformations. The lack of an ordered structure does not imply a lack of function; rather, such regions often participate in interactions with other molecules and become specifically ordered upon complex formation. Disordered regions often contain sites of post-translational modifications (e.g., phosphorylation), which can be predicted using web servers such as DISPHOS [9]. Disordered regions can also be predicted by a number of methods available as web servers, for example, a disorder-predicting meta-server [25]. It has been noted that AS – in particular, intron retention – often leads to the appearance of regions of disorder [26]. However, the prediction of conformational changes on interaction with other molecules is beyond the limits of simple modeling tools, and for the purpose of this protocol disordered regions will be regarded as "unmodelable." Long disordered regions may exhibit sequence biases that often cause the target sequence to be erroneously aligned to unrelated sequences, and therefore they should be excluded from further steps of modeling analyses. Although short disordered regions (e.g., flexible loops) may be

retained for modeling, it should be remembered that the modeling programs will treat them as potentially rigid, and this may lead to various artifacts.

54.2.3 Predicting Transmembrane Helices, Coiled-Coils, and Repeats

In principle, protein structure modeling is applicable to all regions of sequence predicted to be ordered. However, there are certain types of protein structure that require special treatment due to, for example, biases in sequence composition or repetitive character. In particular, sequence analysis prior to modeling should include the detection of transmembrane regions (e.g., by OCTOPUS [10]), coiled coils (e.g., by PCOILS [11]), and repeated segments (e.g., by REPPER [12]). These regions should be removed from further steps of modeling analyses, or they should be modeled independently from other segments of protein sequence.

54.2.4 Protein Fold Recognition

The inference of homology via the detection of sequence similarity is a key to template-based protein structure prediction. The sequences of globular domains to be modeled should be submitted for the "fold recognition" procedure; an example is the detection of similarity of the target sequence to proteins with structures available in the PDB – that is, potential templates for modeling. A number of different fold-recognition methods exist, and it has been established that the best results are achieved if a consensus approach is used. Several fold-recognition methods can be queried simultaneously via one of the available meta-servers, such as the GENESILICO META-SERVER [3]. As a result of the fold-recognition procedure, the user obtains a series of alignments between the target sequence and sequences of potential templates, as well as a consensus prediction made with the PCONS method [13], all with scores indicating the likelihood of correct prediction.

A procedure recommended for inexperienced investigators is to check whether the scores of the HHSEARCH [14] and PCONS methods exceed the threshold of 95% reliability (threshold values can be taken from the results of the Livebench experiment [27]). If the first match reported by HHSEARCH exhibits a significant score, then the corresponding template and alignment can be taken as a working model for further analyses. The first prediction made by PCONS can be used if HHSEARCH fails to report a well-scored template. If neither of these methods reports a confident prediction, this usually indicates either the absence of a suitable template and/or a difficult case of modeling (either with template-bases or template-free tools) that requires the intervention of an expert.

54.2.5 Target–Template Alignment

Fold-recognition methods often make errors in alignments, especially if they report matches to remotely related templates. Therefore, the target–template alignment should be analyzed for potential errors, such as the placement of insertions and deletions in the protein core, or mismatches between functionally important residues in the target and homologous residues in the template. Errors in the alignment must be corrected before the target is modeled; for example, the sites of insertions and deletions should usually be shifted to surface-exposed regions such as loops. Target–template alignments can be edited in programs for protein structure visualization such as SWISS-PDB-VIEWER [15], or in alignment editors such as BioEdit or Jalview [16].

At this point, the investigator must critically analyze (in the light of the results obtained thus far) how the change in protein sequence due to AS may affect the structure of the target protein.

In the case of a *deletion*, the following issues must be addressed:

- What is the nature of the region to be deleted? Is it a part of the ordered domain, or a disordered region? The deletion of entire domains or disordered regions usually does not affect the structure of the remaining domains. However, the deletion of a sequence that forms the hydrophobic core of a globular domain usually results in severe structural changes that cannot be reproduced by the template-based modeling procedure. The template-based modeling of a deletion in the core may result in a structure with an artificial cavity which, in reality, would collapse.
- What is the distance between the amino acid residues flanking the deleted regions in the template structure? Is it possible to "close" the protein backbone by removing a segment and simply sealing the ends, without causing major changes in the structure of the whole domain? If the ends of deleted region are located too far from each other, the procedure of "ligation" may either artificially disrupt the flanking elements of secondary structure, or force the modeling program to thread the resulting linker via the protein core, thereby creating an artificial knot.

In the cases of *substitution* and *insertion*, the following issues must be addressed:

- What is the nature of the inserted or substituted region? If it constitutes a separate ordered domain, it should be modeled separately.
- Is the given region ordered or disordered, and does it have a predicted secondary structure? If it is predicted to be disordered, it may be modeled as a loop extruding from the protein surface (*Note*: the structure modeling programs are not designed to model the dynamics of loops). If the inserted sequence is predicted to be ordered and to possess secondary structure, it may be modeled as a loop and then locally remodeled *de novo*, using methods such as ROSETTA.

54.2.6 Template-Based Modeling

The refined target–template alignment and the structure of the template constitute a minimal input to most of the modeling programs. Most such programs include a method for the rudimentary optimization of model geometry, which allows for the creation of structures without severe steric clashes. Currently, many programs are available with which to perform model building. For the inexperienced user, a web-based program such as SWISS-MODEL is recommended, as this can take as an input the "project" files prepared in a molecular viewer SWISS-PDB-VIEWER. MODELLER is another commonly used program for template-based modeling; this allows the user to include additional restraints, for example, to enforce the formation of a particular secondary structure or the distances between selected residues.

54.2.7 Model Quality Assessment

The comparative modeling approach can generate erroneous models, if based on incorrect templates or alignments, and therefore the critical assessment of model accuracy is an essential step of structure prediction. As noted above, the modeling procedure involves copying as many features from the template structure as reasonably possible, and then subjects the model to geometry optimization. Thus, methods that are typically used for the evaluation of quality for crystal structures (e.g., analysis of the Ramachandran plot) are not appropriate for analyzing the quality of theoretical models. Several so-called Model Quality Assessment Programs (MQAPs) have been developed to identify potential errors in theoretical models, and most of

these rely on empirical potentials of mean force derived from statistical analyses of features in known protein structures. For inexperienced users, it is recommended that the models are evaluated with programs which are available as web servers and which provide predictions for both global and local model quality; examples include PROQ and METAMQAP. The evaluated models can be analyzed with a molecular viewer such as SWISS-PDB-VIEWER or RASMOL [17], which can visualize the predicted quality by coloring individual residues according to their score. Regions that are predicted as likely to be erroneous may be subjected to remodeling, for example, by modification of the target–template alignment (see Section 54.2.5) or by using "*de novo*" modeling methods such as ROSETTA to "refold" the suspicious segment. In the case of a low global score, alternative templates may be selected for modeling (see Section 54.2.4).

54.2.8
Is the Same Possible for RNA 3-D Structure Prediction?

The field of RNA 3-D structure prediction lags considerably behind the methods for protein structure modeling. There exist manual structure modeling methods such as S2S [18], but only recently have the first fully automated methods been developed for template-based or *de novo* modeling. Freely available methods for the comparative modeling of RNA 3-D structures include MODERNA (as developed in the present authors' laboratory [19]). *De novo* folding methods include FARNA, a part of the standalone ROSETTA package [28] and iFOLDRNA, which is available as a server [20].

54.3
Example Experiment

Here, a demonstration is provided of how to predict the structure of an alternatively spliced variant of human low-molecular-weight phosphotyrosine phosphatase (LMPTP). The gene of LMPTP has seven exons [29], with exons 3 and 4 being alternatively utilized and giving rise to two isoenzymes, LMPTP-A and LMPTP-B [30], the structures of which have been resolved [31]. The altered segment of the protein surrounds the active site and is responsible for the substrate specificity; therefore, LMPTP-A and LMPTP-B are thought to act on distinct substrates. Subsequently a new isoform, LMPTP-C, has been identified [32] in which neither exon 3 nor exon 4 are used, and which results in a shortened variant with amino acids 40–71 missing. Although LMPTP-C was found to be inactive, it is nevertheless expressed in relatively high concentrations and shares similar epitopes with the full-length variants, which suggests that they may have similar structures. The results of experimental studies suggest that LMPTP-C may act as an antagonist of LMPTP-A and LMPTP-B, although no structural or physiological mechanism of this phenomenon has been proposed.

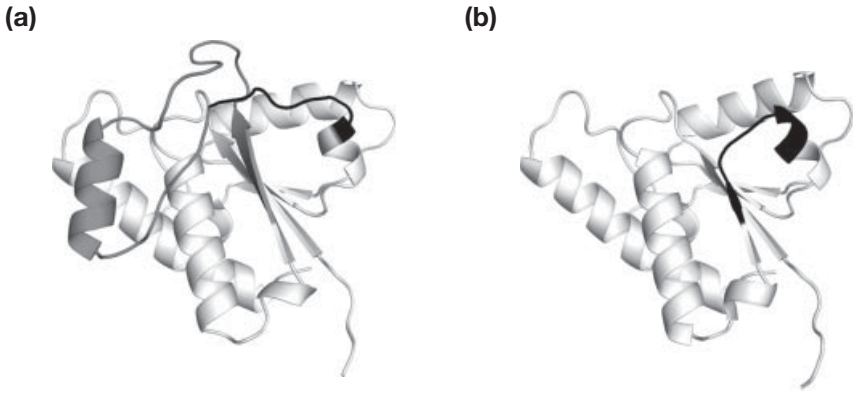


Fig. 54.1 (a) Crystal structure of the full-length human low-molecular-weight phosphotyrosine phosphatase (LMPTP-A), PDB code 5pnt; (b) Predicted structure of the shortened alternatively spliced variant LMPTP-C). The region deleted by the alternative splicing event is colored gray; the region of local structural change in LMPTP-C is colored black.

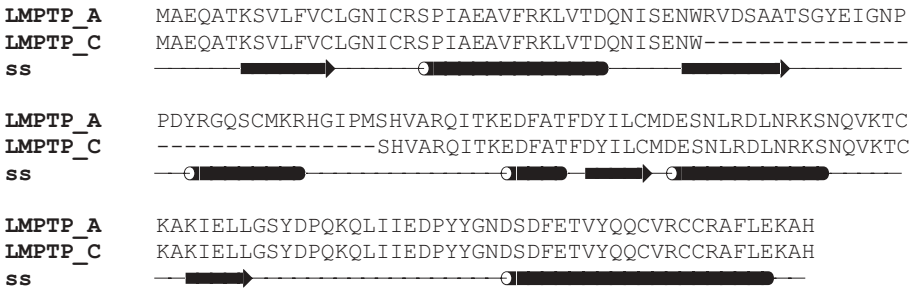


Fig. 54.2 Sequence alignment of full-length human low-molecular-weight phosphotyrosine phosphatase (LMPTP-A) and its truncated variant (LMPTP-C) reported by the fold-recognition method PHYRE, used to model the structure of the latter. ss = secondary structure of LMPTP-A derived from crystal structure. The cylinders represent α-helices; the arrows represent β-strands.

Sequence analyses of LMPTP-C performed using the GENESILICO META-SERVER revealed a single-domain with no predicted disordered regions. The structure of the splicing variant LMPTP-A (PDB code 5pnt; Figure 54.1a) was proposed as the best template for structure prediction by most of the fold-recognition methods used. The alignment reported by the PHYRE method [21] (Figure 54.2) was selected arbitrarily to create a project with the aid of the SWISS-PDB VIEWER molecular viewer. Mapping of the sequence on the template structure revealed that the deletion in LMPTP-C spans one β-strand and one loop of the substrate-binding site in LMPTP-A. The structural elements missing from LMPTP-C are located on the surface of the protein, and appear not to disrupt the hydrophobic core of the protein, which is consistent with the observation that the LMPTP-C variant is able to fold. Although the distance between the ends of deleted regions was quite large (20 Å), an inspection of the target–template alignment in SWISS-PDB-VIEWER revealed that a slight shift in the alignment can bring together the terminal residues encoded by exon 2 and exon 5, to close a gap without major structural alterations other than omission of the deleted segment. The modeling project was saved and submitted to comparative modeling by the SWISSMODEL server. The predicted structure (Figure 54.1b) was evaluated by MetaMQAPII as reasonably good (predicted RMSD to the unknown real structure 2.75 Å). Most importantly, no major errors are predicted to exist in the area of the deletion – that is, in the remodeled fragment itself – as well as in the unchanged regions with which it interacts.

Table 54.1 Programs proposed by the authors to be used during the course of the modeling exercise. This table and its URLs are available online www.wiley-vch.de/home/splicing

Program	URL	Reference
GENESILICO META-SERVER	https://genesilico.pl/meta2/	[3]
MODELLER	http://salilab.org/modeller/	[4]
SWISS-MODEL	http://swissmodel.expasy.org/	[5]
PROQ	http://www.sbc.su.se/~bjornw/ProQ/ProQ.cgi	[6]
METAMQAP	https://genesilico.pl/toolkit/unimod?method=MetaMQAPII	[7]
ROSETTA	http://www.rosettacommons.org/	[8]
DISPHOS	http://core.ist.temple.edu/pred/	[9]
OCTOPUS	http://octopus.cbr.su.se/index.php?about=OCTOPUS	[10]
PCOILS	http://toolkit.tuebingen.mpg.de/pcoils	[11]
REPPER	http://toolkit.tuebingen.mpg.de/repper	[12]
PCONS	http://pcons.net/	[13]
HHSEARCH	http://toolkit.tuebingen.mpg.de/hhpred	[14]
BioEdit	http://www.mbio.ncsu.edu/BioEdit/	Ibis Biosciences
SWISS-PDB-VIEWER	http://spdbv.vital-it.ch/	[15]
Jalview	http://www.jalview.org/	[16]
RASMOL	http://rasmol.org/	[17]
S2S	http://www.bioinformatics.org/S2S/	[18]
MODERNA	http://genesilico.pl/moderna/	[19]
iFOLDRNA	http://troll.med.unc.edu/ifoldrna/	[20]
PHYRE	http://www.sbg.bio.ic.ac.uk/phyre/	[21]
DPANN	http://nihserver.mbi.ucla.edu/cgi-bin/DPANN/DPANN-Interface.cgi	[22]

54.4 Troubleshooting (see Table 54.1)

Problem	Reason + Solution
Cannot find a modeling template, target–template alignments returned by fold-recognition methods exhibit poor scores	1. There is no homologous template in the current database. Template-free modeling can be attempted (but the chances of successful modeling with this approach are low). 2. The existing templates are too diverged to be directly detectable by current methods. Try searching for additional homologs among other protein families, and then attempt protein fold-recognition using a multiple sequence alignment as a query
The fold is predicted confidently, but the target–template alignment contains multiple deletions and insertions in the core. It also appears that, in some regions, the sequences are mismatched	Protein fold-recognition methods are optimized to identify correct folds, rather than to generate optimal alignments. Use methods that refine target–template alignments, for example, DPANN [22] or the Frankenstein's monster approach [33]
The global score of the model is low	1. The modeling program introduced a critical artifact. Compare the model to the template. If some drastic change occurred (e.g., the structure of the model appears disrupted or knotted), try to modify the target–template alignment in the vicinity of the suspected modification and repeat the modeling procedure (repeat steps 54.2.6 and 54.2.7). 2. A wrong template has been selected. Try another template (exclude the previous template and repeat steps 54.2.4–54.2.7) 3. The target–template alignment contains too many errors. Try a different alignment (e.g., proposed by a different server) to the same template structure or to its homolog and repeat the modeling procedure (repeat steps 54.2.5–54.2.7).

Acknowledgments

These studies were supported financially by grants of the 6th and 7th Framework Program of the EU (LSHG-CT-2005-518238 and 229676), and by the Polish Ministry of Science and Higher Education (grants PBZ-MNiI-2/1/2005, 188/N-DFG/2008/0, N N301 297337 and POIG.02.03.00-00-003/09).

References

1 Anfinsen, C.B. (1973) *Science*, **181**, 223–230.
2 Chothia, C. and Lesk, A.M. (1986) *EMBO J.*, **5**, 823–826.
3 Kurowski, M.A. and Bujnicki, J.M. (2003) *Nucleic Acids Res.*, **31**, 3305–3307.
4 Sali, A. and Blundell, T.L. (1993) *J. Mol. Biol.*, **234**, 779–815.
5 Schwede, T., Kopp, J., Guex, N., and Peitsch, M.C. (2003) *Nucleic Acids Res.*, **31**, 3381–3385.
6 Wallner, B. and Elofsson, A. (2003) *Protein Sci.*, **12**, 1073–1086.
7 Pawlowski, M., Gajda, M.J., Matlak, R., and Bujnicki, J.M. (2008) *BMC Bioinformatics*, **9**, 403.

8 Simons, K.T., Kooperberg, C., Huang, E., and Baker, D. (1997) *J. Mol. Biol.*, **268**, 209–225.
9 Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., and Dunker, A.K. (2004) *Nucleic Acids Res.*, **32**, 1037–1049.
10 Viklund, H. and Elofsson, A. (2008) *Bioinformatics*, **24**, 1662–1668.
11 Gruber, M., Soding, J., and Lupas, A.N. (2006) *J. Struct. Biol.*, **155**, 140–145.
12 Gruber, M., Soding, J., and Lupas, A.N. (2005) *Nucleic Acids Res.*, **33**, W239–W243.
13 Lundstrom, J., Rychlewski, L., Bujnicki, J., and Elofsson, A. (2001) *Protein Sci.*, **10**, 2354–2362.
14 Hildebrand, A., Remmert, M., Biegert, A., and Söding, J. (2009) *Proteins*, **77** (Suppl.), 128–132.
15 Guex, N. and Peitsch, M.C. (1997) *Electrophoresis*, **18**, 2714–2723.
16 Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009) *Bioinformatics*, **25**, 1189–1191.
17 Sayle, R.A. and Milner-White, E.J. (1885) *Trends Biochem. Sci.*, **20**, 374.
18 Jossinet, F. and Westhof, E. (2005) *Bioinformatics*, **21**, 3320–3321.
19 Musielak, M., Rother, K., Puton, T., and Bujnicki, J.M. (2011) *Nucleic Acids Res.*, **39**, 4007–4022. Available at: http://genesilico.pl/moderna/.
20 Sharma, S., Ding, F., and Dokholyan, N.V. (2008) *Bioinformatics*, **24**, 1951–1952.
21 Bennett-Lovsey, R.M., Herbert, A.D., Sternberg, M.J., and Kelley, L.A. (2007) *Proteins*, **70**, 611–625.
22 Reinhardt, A. and Eisenberg, D. (2004) *Proteins*, **56**, 528–538.
23 Levitt, M. (2009) *Proc. Natl Acad. Sci. USA*, **106**, 11079–11084.
24 Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L., and Bateman, A. (2008) *Nucleic Acids Res.*, **36**, D281–D288.
25 Ishida, T. and Kinoshita, K. (2008) *Bioinformatics*, **24**, 1344–1348.
26 Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M., LeGall, T., Obradovic, Z., and Dunker, A.K. (2006) *Proc. Natl Acad. Sci. USA*, **103**, 8390–8395.
27 Rychlewski, L. and Fischer, D. (2005) *Protein Sci.*, **14**, 240–245.
28 Das, R. and Baker, D. (2007) *Proc. Natl Acad. Sci. USA*, **104**, 14664–14669.
29 Bryson, G.L., Massa, H., Trask, B.J., and Van Etten, R.L. (1995) *Genomics*, **30**, 133–140.
30 Wo, Y.Y., McCormack, A.L., Shabanowitz, J., Hunt, D.F., Davis, J.P., Mitchell, G.L., and Van Etten, R.L. (1992) *J. Biol. Chem.*, **267**, 10856–10865.
31 Zhang, M., Stauffacher, C.V., Lin, D., and Van Etten, R.L. (1998) *J. Biol. Chem.*, **273**, 21714–21720.
32 Tailor, P., Gilman, J., Williams, S., and Mustelin, T. (1999) *Eur. J. Biochem.*, **262**, 277–282.
33 Kosinski, J., Gajda, M.J., Cymerman, I.A., Kurowski, M.A., Pawlowski, M., Boniecki, M., Obarska, A., Papaj, G., Sroczynska-Obuchowicz, P., Tkaczuk, K.L., Sniezynska, P., Sasin, J.M., Augustyn, A., Bujnicki, J.M., and Feder, M. (2005) *Proteins*, **61** (Suppl. 7), 106–113.